Naval Aerospace Medical Research Laboratory

NAMRL Special Report 98-1

PSYCHOMETRIC EQUIVALENCY ISSUES FOR THE APEX SYSTEM

D. J. Blower

Naval Aerospace Medical Research Laboratory
51 Hovey Road
Pensacola, Florida 32508-1046

Reviewed and approved 23 Apr 98

L. H. FRANK, CAPT, MSC USN
Commanding Officer

This research was sponsored by the Naval Medical Research and Development Command under work unit 64771NMM33P30.001 7802 DN243515.

The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government.

Volunteer subjects were recruited, evaluated, and employed in accordance with the procedures specified in the Department of Defense Directive 3216.2 and Secretary of the Navy Instruction 3900.39 series. These instructions are based upon voluntary informed consent and meet or exceed the provisions of prevailing national and international guidelines.

Trade names of materials and/or products of commercial or nongovernment organizations are cited as needed for precision. These citations do not constitute official endorsement or approval of the use of such commercial materials and/or products.

NAVAL AEROSPACE MEDICAL RESEARCH LABORATORY
51 HOVEY ROAD, PENSACOLA, FL 32508-1046

NAMRL Special Report 98-1

PSYCHOMETRIC EQUIVALENCY ISSUES FOR THE APEX SYSTEM

D. J. Blower

ABSTRACT

This report discusses some of the psychometric properties of the Aviation Selection Test Battery (ASTB) when administered in the standard paper and pencil format as compared to an experimental version administered on the computer. Structural Equation Modeling is employed as the analytical framework to address certain questions of psychometric equivalency. The two experimental conditions are compared in their effects on the internal reliability and validity on eight subtests of the ASTB, the means of these scores, and finally on the structural coefficients when these scores are used in a predictive validity setting. The two modes of presentation for the ASTB, the traditional paper and pencil version and the experimental computer version, do not seem to differ in many psychometric properties. These results, then, are important in deflecting criticism concerning the impact of this new presentation medium on ASTB test scores.

Acknowledgments

I would like to acknowledge the support of LCDR (sel) Sean Biggerstaff, Head of the Aviation Selection Division at NAMRL when this work was done, and my colleagues in the Aviation Selection Division, Mr. Allen Chapman and Ms. Claire Portman. Claire was especially helpful in providing the database utilized in this analysis.

Introduction

The Automated Pilot Examination (APEX) System is an ongoing research effort designed to improve upon a selection test for individuals interested in becoming Naval and Marine Corps aviators. The Aviation Selection Test Battery (ASTB) is the Navy's current selection instrument and it is used along with other medical criteria to determine acceptance into flight training. The ASTB is a conventional paper and pencil test which attempts to measure certain cognitive skills thought to be predictive of success at least through primary flight training.

The first iteration of APEX simply translated the ASTB paper and pencil test into a test presented on a computer. The only differences a test taker would notice are the obvious ones; the questions are presented on a computer monitor, answers are selected via mouse input, *etc.* Otherwise, it is a verbatim copy of the paper and pencil ASTB. The most current version of APEX has an improved user interface different from the one used to gather the data reported here.

Despite this exact similarity in test content, one could question whether the mode of presentation, *i.e.*, paper and pencil vs. computer, does not somehow induce subtle, or perhaps not so subtle, changes in the character of the test; changes that would cause a test taker to score differently on the two modes of presentation. To answer questions of this sort, this report addresses the issue of psychometric equivalency between the computer mode of presentation of the APEX system *vis-à-vis* the standard mode of the paper and pencil version.

The major problem in the statistical analysis of selection data is the fallible nature of test instruments. Tests that purport to measure some underlying skill or capability always have some error attached to this assessment of the true skill. Structural equation modeling (Hayduk (1), Loehlin (2), and Bollen (3)) seems to be the preferred method of dealing with fallible test instruments. In this report, a quantitative approach to psychometric equivalency issues is therefore defined through structural equation modeling. The software package LISREL 8 (Jöreskog and Sörbom (4)) is employed to find estimates, standard errors, and t-values for the elements of the parameter matrices that appear in the various structural equation models. LISREL also provides chi-square values to ascertain the goodness of fit of the various proposed models.

This report looks at psychometric equivalency from three broad perspectives. Did the change in the mode of presentation affect

[1] the observed variances and covariances of the tests for the two groups?
[2] the means of the two groups?
[3] the predictive validity of the ASTB on criteria of flight training success for the two groups?

Experimental Design

This section gives a brief overview of the experimental design sufficient for the purposes of this report. For further details see the accompanying report by Biggerstaff, Portman, Blower, and Chapman (5). 82 subjects participated in the study with 42 subjects taking the paper and pencil ASTB and 40 subjects taking the ASTB under the new computerized mode of presentation. All subjects were Navy and Marine Corps officers awaiting ground school instruction prior to primary flight training. These tests were taken under laboratory control. All subjects (except one) had taken a different form of the paper and pencil ASTB some time earlier at recruiting stations, the Naval Academy, during ROTC, *etc.* During the experiment, in addition to the official ASTB, subjects took an alternate ASTB in the same mode of presentation as they took the official ASTB.

The ASTB consists of five subtests, abbreviated for future reference as follows:

[1] mathematics and verbal subtest (MVT)
[2] mechanical comprehension subtest (MCT)
[3] spatial apperception subtest (SAT)
[4] aviation/nautical interest subtest (ANI)
[5] biographical inventory subtest (BI)

Only the MVT, MCT, SAT, and ANI subtests are analyzed in this report; the BI is excluded. The version of the four subtests taken previous to the laboratory administered tests have a "P" appended, while the alternate versions have an "A" appended. Thus, there are 12 data points recorded for each subject (repeated measures design) in each of the two groups, the subtest score on MVT, MCT, SAT, ANI, MVTP, MCTP, SATP, ANIP, MVTA, MCTA, SATA, and ANIA. Of these 12 scores, only 8 will be analyzed here, *viz.*, the previous official ASTB consisting of MVTP, MCTP, SATP, and ANIP, and the laboratory administered official ASTB consisting of MVT, MCT, SAT and ANI. The primary emphasis will be on comparing the group who took the test in the usual paper and pencil format with the group who took the test in the new computerized format. Subjects in both groups took the previous version as a paper and pencil test.

The Observed Data

Mathematically, structural equation modeling relies on the assumption of multivariate normality for the observed scores on the tests. Therefore, sample means and the sample variance-covariance matrices are sufficient statistics and serve as the data to be analyzed. The first part of the analysis looks at the internal reliability and validity of the eight tests, four subtests taken in the laboratory setting and four subtests taken previously. Tables 1 and 2 show the observed sample variance-covariance matrices for the $N = 42$ subjects in the paper and pencil group and the $N = 40$ subjects in the computer group, respectively.

Table 1: The sample variance-covariance matrix $S^{(1)}$ for $N = 42$ subjects in the paper and pencil group.

	MVT	MCT	SAT	ANI	MVTP	MCTP	SATP	ANIP
MVT	34.611							
MCT	15.169	15.051						
SAT	4.115	9.348	36.906					
ANI	8.585	8.146	1.537	17.984				
MVTP	20.751	8.840	-5.302	.334	30.662			
MCTP	7.580	5.542	1.805	3.718	7.123	10.730		
SATP	-3.257	1.989	24.102	-2.194	-8.434	-1.573	27.476	
ANIP	-0.125	0.150	-1.050	4.150	-0.900	2.900	1.175	11.650

There are 36 values in each of these two matrices. The notation given to the sample variance-covariance matrix is $S^{(1)}$ for the paper and pencil group and $S^{(2)}$ for the computer group.

After discussing whether there have been any changes in the psychometric properties of the variance-covariance

Table 2: The sample variance-covariance matrix $S^{(2)}$ for $N = 40$ subjects in the computer group.

	MVT	MCT	SAT	ANI	MVTP	MCTP	SATP	ANIP
MVT	26.964							
MCT	9.979	17.230						
SAT	0.587	4.667	27.128					
ANI	3.336	7.301	3.135	13.717				
MVTP	17.605	6.259	1.664	1.131	17.579			
MCTP	11.105	10.993	2.289	8.743	8.079	18.487		
SATP	-1.754	2.445	20.350	2.663	1.221	3.253	32.182	
ANIP	5.918	8.322	3.576	10.206	2.200	8.847	3.367	12.715

matrices for the two groups, the means on the eight tests are analyzed for any differences. Table 3 shows the means, standard deviations, and sample size for the two modes of presentation.

Table 3: Means, standard deviations, and sample sizes for the two groups over the eight tests.

Test	Paper and Pencil			Computer		
	Mean	SD	N	Mean	SD	N
MVT	26.79	5.88	42	27.90	5.19	40
MCT	22.21	3.88	42	21.53	4.15	40
SAT	25.86	6.08	42	27.48	5.21	40
ANI	19.33	4.24	42	19.77	3.70	40
MVTP	27.71	5.54	41	27.90	4.19	40
MCTP	22.66	3.28	41	21.77	4.30	40
SATP	27.22	5.24	41	26.65	5.67	40
ANIP	18.00	3.41	41	19.45	3.57	40

The third check on a possible psychometric disturbance due to changing the way the tests are presented involves adding a criterion variable. The data for this analysis is the same variance-covariance matrix as given above in Tables 1 and 2 with the addition of the new information about the variance of the criterion variable and its covariance with all eight tests. These two matrices would then consist of 45 values. Only the 9 new values of the variance-covariance matrix are shown in Table 4 below. These augmented matrices for the two groups serve then as the observed data for the third part of the analysis.

LISREL Notation

We give here a brief summary of the notation used in LISREL and in the subsequent sections of this report. Unfortunately, a rather lengthy prelude is necessary in order to introduce the various elements in the LISREL model. This is important because these elements operationally define what is meant by psychometric equivalency.

The LISREL model is broken down into two components: (1) the measurement model, and (2) the structural equation model. All that is needed for the first analysis is the LISREL measurement model. Traditionally, the

Table 4: The variance of the criterion variable and its covariances with the eight tests for the two modes of presentation.

Test	Paper and Pencil	Computer
Criterion	51.313	59.553
MVT	20.199	14.466
MCT	15.302	3.986
SAT	20.020	15.462
ANI	12.863	5.937
MVTP	11.870	12.071
MCTP	8.270	9.896
SATP	13.390	9.395
ANIP	1.267	6.151

LISREL notation has used Greek letters for the parameter matrices that appear in the model. We will not depart from this convention which has become fairly well standardized in the literature. The measurement model is defined as,

$$x = \Lambda_x \xi + \delta \tag{1}$$

where x is a column vector of the eight test scores (MVT, MCT, SAT, ANI, MVTP, MCTP, SATP, and ANIP). x in the measurement model represents deviations about the mean, so the numbers in x are the actual test scores with their means subtracted. Λ_x is the matrix of factor loadings of the eight tests on any underlying latent variables or factors. ξ is the column vector of the latent variables or factors, while δ is the column vector of measurement errors in the model.

It was mentioned in the previous section that the sample variances and covariances among the eight tests form the data for this analysis. The theoretical variance-covariance matrix that arises for the measurement model in Equation (1) is called Σ.

$$\Sigma = E[xx^T] \tag{2}$$

$$E[xx^T] = E[(\Lambda_x \xi + \delta)(\Lambda_x \xi + \delta)^T] \tag{3}$$

$$\Sigma = \Lambda_x \Phi \Lambda_x^T + \Theta_\delta \tag{4}$$

An adequate fit of the model-implied variance-covariance matrix Σ to the sample variance-covariance matrix S is the objective of the numerical routines in LISREL. This is accomplished by adjusting the values that can appear in Λ_x, Φ, and Θ_δ. Λ_x is the same matrix of factor loadings defined in Equation (1). Φ is the matrix of the variances and covariances among the factors, ξ_i, while Θ_δ is the matrix of the variances and covariances among the δ_i. In passing, we note that Equations (1) and (4) are equivalent to traditional factor analysis models.

Λ_x, Φ, and Θ_δ are called parameter matrices and our attention will be focused on these three matrices throughout the rest of this report. We are especially interested in whether any or all of these parameter matrices differ as the result of the mode of presentation. A superscript (1) or (2) will be appended to these parameter matrices to indicate just which group is being referenced. Group 1 is the paper and pencil mode of presentation

while Group 2 is the computer mode of presentation. *Assessing the differences (or similarities) among the elements of these three parameter matrices is the operational definition of determining psychometric equivalency between the two groups.*

As a final closure to the notation, we mention the size of the various vectors and matrices in the measurement model. $\boldsymbol{x}$ is a $q \times 1$ column vector of the eight test score deviations, therefore, for this analysis, $\boldsymbol{x}$ is 8×1. $\boldsymbol{\xi}$ is a $n \times 1$ column vector of the latent variables or factors. n will vary as we consider different good and bad measurement models. Jumping ahead a bit, we will find that $n = 4$ results in an acceptable model. Because the δ_i are the errors attached to each test, $\boldsymbol{\delta}$ is also a $q \times 1$ or 8×1 column vector.

Now we state the sizes of the three parameter matrices. $\boldsymbol{\Lambda_x}$ is a $q \times n$ matrix so, for example as shown later for one of the good models, $\boldsymbol{\Lambda_x}$ is an 8×4 matrix of factor loadings of eight tests on four factors. $\boldsymbol{\Phi}$ is $n \times n$ and is therefore a 4×4 matrix of factor variances and covariances for the good model. Because this matrix is symmetric, only the $[n \times (n+1)]/2 = 10$ diagonal and lower subdiagonal elements are needed. Lastly, $\boldsymbol{\Theta_\delta}$ is a $q \times q$ matrix and in its totality would be an 8×8 matrix. However, most often we deal only with the diagonal elements of $\boldsymbol{\Theta_\delta}$ so we are reduced to only $q = 8$ elements. Sometimes to find a good fit of $\boldsymbol{\Sigma}$ to S, it is necessary to estimate as well one or two off-diagonal elements within $\boldsymbol{\Theta_\delta}$.

This section can be amplified and made somewhat clearer by consideration of Figure 1. This figure sketches a LISREL measurement model with eight tests ($q = 8$) and four underlying factors ($n = 4$). The eight tests are written within the rectangular boxes and the four underlying latent variables are indicated by the circles. The tests are the $\boldsymbol{x}$ and the latent variables are the $\boldsymbol{\xi}$. These latent variables are supposed to represent some true underlying cognitive skills labeled (1) Quantitative/Verbal, (2) Mechanical Comprehension, (3) Spatial Apperception, and (4) Aviation/Nautical Interest. Arrows connect these underlying variables to the corresponding tests that they supposedly determine. The loadings of the tests on the factors are written over the connecting arrows as λ_{ij}. These are the $q \times n$ elements of $\boldsymbol{\Lambda_x}$. If there is no arrow connecting a factor (ξ_i) with a test, then its corresponding $\lambda_{ij} = 0$. The measurement errors are shown by a second set of connecting arrows leading from the δ_i to the tests. The final set of arrows are double-headed arrows connecting the four underlying latent variables. Next to these are written the elements of $\boldsymbol{\Phi}$, ϕ_{21}, ϕ_{31}, ϕ_{32} and so on.

Search History for Acceptable Models

The primary focus of structural equation modeling is not so much upon finding *one* model, but rather emphasizes finding the rough dividing line between a large class of good models and the much larger class of bad models. In this section we illustrate this philosophy by listing some examples of bad models for the data in this study, and then show that there are also many acceptable models. Without further research and data, one cannot categorically latch onto any one of these acceptable models to the exclusion of the other acceptable models. Science is a converging and iterative process, and in these early research stages of APEX it would be unwise to place all of our eggs in one basket.

LISREL 8 must be run many, many times in a search for the class of good models. This section gives an abbreviated synopsis of such a search history. It is easy to find obviously bad models and these are discussed first. Then we discuss models that are on the verge of transitioning from the class of bad models to the class of good models. Finally, we take a look at some of the acceptable models. The estimates of the parameter matrices for one of these good models is elaborated on in the next section.

Table 5 lists eight bad models out of a vast array of potential candidates. These bad models all exhibit

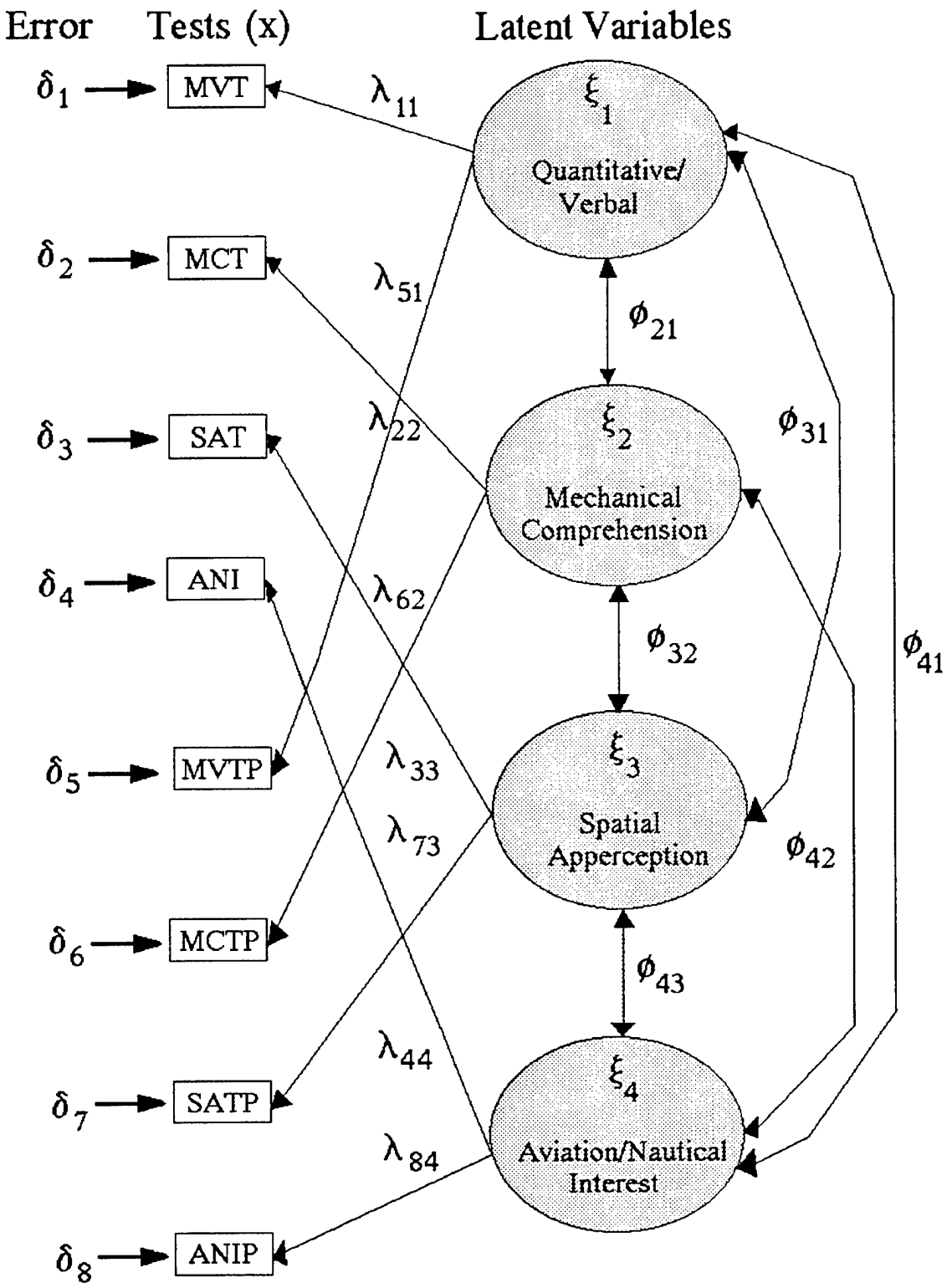

Figure 1: A sketch of a LISREL measurement model with eight tests and four underlying factors.

excessive misfit of the theoretical Σ implied by Equation (4) to the data in S. This is evidenced by the large χ^2 values relative to the degrees of freedom (df). The table also shows the interplay among the parameter matrices between the two groups as the equality constraints are relaxed.

Table 5: Initial phase in search of good models. The following eight models are all bad models. The model-implied Σs and the acceptability of the fit from the model to the data are all quite unacceptable as shown by the large χ^2 values.

Model	χ^2	df	p-value	Decision
A	213.62	63	$< .0001$	Reject
B	205.35	55	$< .0001$	Reject
C	204.76	54	$< .0001$	Reject
D	188.98	47	$< .0001$	Reject
E	194.01	61	$< .0001$	Reject
F	179.00	53	$< .0001$	Reject
G	174.57	50	$< .0001$	Reject
H	174.05	48	$< .0001$	Reject

For example, Model A assumes a factor structure with only one latent variable, ξ_1. Perhaps all of the tests simply load on one general intelligence factor, the g factor. The factor loadings for the paper and pencil group are set at,

$$\mathbf{\Lambda_x}^{(1)} = \begin{bmatrix} 1.0 \\ 0.9 \\ 0.8 \\ 0.5 \\ 1.0 \\ 0.9 \\ 0.8 \\ 0.5 \end{bmatrix} \tag{5}$$

Strong equality constraints are postulated among the three parameter matrices such that,

$$\mathbf{\Lambda_x}^{(1)} = \mathbf{\Lambda_x}^{(2)} \tag{6}$$

$$\mathbf{\Phi}^{(1)} = \mathbf{\Phi}^{(2)} \tag{7}$$

$$\mathbf{\Theta_\delta}^{(1)} = \mathbf{\Theta_\delta}^{(2)} \tag{8}$$

Such equality constraints as represented by Equations (6)–(8) are the definition of very strong psychometric equivalency between the two groups. There are a total of 72 df, 36 df for each of the two sample variance-covariance matrices. 9 df are consumed in (1) the estimation of $\phi_{11}^{(1)}$, the variance of the one latent variable for the paper and pencil group, and in (2) the estimation of $\theta_{11}^{(1)}$ through $\theta_{88}^{(1)}$, the variances of the specific and measurement error for the paper and pencil group. $\lambda_{11}^{(1)}$ through $\lambda_{81}^{(1)}$ are not estimated, but fixed by theoretical assumptions as shown above. Therefore, no df are lost when elements in a parameter matrix are

specified by theoretical concerns. That this is a bad model is highlighted by,

$$\chi^2(63\ df) = 213.62 \qquad p < .0001 \tag{9}$$

We cannot seriously entertain a model with only one underlying latent variable for all eight tests with equality for all of the parameter matrices.

Model B attempts to improve upon this model by relaxing the equality constraint between the error variances of the two groups so that now only,

$$\Lambda_x{}^{(1)} = \Lambda_x{}^{(2)} \tag{10}$$

$$\Phi^{(1)} = \Phi^{(2)} \tag{11}$$

and $\Theta_\delta^{(2)}$ can be freely estimated independent of the estimates for $\Theta_\delta^{(1)}$. Because there are eight more estimates, eight more df are lost, and the χ^2 test will use 55 df. But this relaxation did not sufficiently increase the fit to the data, (the two sample variance-covariance matrices), so this model must be rejected as well.

Model C relaxes the equality of the factor variance between the paper and pencil and computer groups, while Model D relaxes the equality of the only parameter matrix left between the two groups, the factor loadings. None of these changes, however, can remove these models from the bad model category because the factor structure is too impoverished.

Models E through H in the bottom part of Table 5 run through the same pattern as Models A through D with the exception that the factor structure is enlarged to *two* underlying latent variables, ξ_1 and ξ_2. The factor loadings for this set of models assumes that the MVT and the MCT both load equally on the first factor while the SAT and the ANI load equally on the second factor. As in the first set of models, it is also assumed that the previous versions of the tests load on the two factors in exactly the same way as the tests taken in the experiment.

$$\Lambda_x{}^{(1)} = \begin{bmatrix} 1.00 & 0.00 \\ 1.00 & 0.00 \\ 0.00 & 1.00 \\ 0.00 & 1.00 \\ 1.00 & 0.00 \\ 1.00 & 0.00 \\ 0.00 & 1.00 \\ 0.00 & 1.00 \end{bmatrix} \tag{12}$$

The same pattern of decreasing χ^2 is observed when the equality between the parameter matrices is allowed to relax, but the factor structure for the second set of models is also not rich enough to adequately account for the data. The χ^2 values are all much too large for the degrees of freedom.

We now transition to an intermediate phase to consider models that remarkably improve upon the bad class of models just discussed, but do not quite reach acceptability. They point in the direction one has to go in order to find the class of good models. Five typical examples from this intermediate phase are presented in Table 6.

Postulating one or two latent variables (factors) underlying the eight tests did not work. We now entertain the reasonable alternative that the four different tests (MVT, MCT, SAT, and ANI) each loads on its own separate

Table 6: Intermediate phase in search for good models. Five model-implied Σs and the acceptability of the fit from the model to the data indicate the direction to search for good models. The usual criterion for accepting a model is $p > .05$.

Model	χ^2	df	p-value	Decision
I	154.43	60	.0001	Reject
J	92.30	54	.0009	Reject
K	83.68	44	.0003	Reject
L	56.92	36	.0150	Reject
M	73.28	53	.0340	Reject

factor. The tests taken previous to the laboratory tests (MVTP, MCTP, SATP, and ANIP) load on the same four factors. Therefore, for all of these models in the intermediate category,

$$\Lambda_x = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{13}$$

In Model I, the very strong psychometric equivalency for all three parameter matrices between the two modes of presentation is established. As we have seen before, this implies,

$$\Lambda_x{}^{(1)} = \Lambda_x{}^{(2)} \tag{14}$$

$$\Phi^{(1)} = \Phi^{(2)} \tag{15}$$

$$\Theta_\delta{}^{(1)} = \Theta_\delta{}^{(2)} \tag{16}$$

Moreover, we also theoretically affirm that the four factors are independent of one another for both groups so that, for example, within $\Phi^{(1)}$, $\phi_{21}^{(1)} = \phi_{31}^{(1)} = \phi_{32}^{(1)} = \phi_{41}^{(1)} = \phi_{42}^{(1)} = \phi_{43}^{(1)} = 0$. 12 df are used up in this model by the estimation of the four factor variances, $\phi_{11}^{(1)}$, $\phi_{22}^{(1)}$, $\phi_{33}^{(1)}$, and $\phi_{44}^{(1)}$, and the eight error variances, $\theta_{11}^{(1)}$ through $\theta_{88}^{(1)}$, that must be estimated. That this set of constraints is too strong is indicated by the large χ^2 in the first row of Table 6 and dictates that Model I must be rejected.

Model J relaxes the assumption that all the factors are uncorrelated. Φ is now estimated for all the covariances that were assumed to be zero in Model I. 6 more df are lost for these estimates. All the other equality constraints between the two groups were retained. With 54 df χ^2 has been reduced to 92.30.

We now try to see if relaxing the equality of $\Phi^{(1)} = \Phi^{(2)}$ results in any improvement. 10 more estimates for $\Phi^{(2)}$ are now required, dropping the df to 44. Unfortunately, the desired drop in χ^2 for Model K is more than matched by the drop in the df, so there is no improvement in this direction. Model L relaxes the equality between

Θ_δ for the two groups. 8 more df are lost on account of the eight new estimates in $\Theta_\delta^{(2)}$. However, we are getting closer to an acceptable model.

In the final model, Model M, the hypothesis is tested that perhaps just some of the factor correlations are zero. Given the content of the tests, it is reasonable to ask whether the Spatial Apperception skill (the underlying latent variable ξ_3) is independent of the other three skills. Therefore, $\phi_{31}^{(1)} = \phi_{32}^{(1)} = \phi_{43}^{(1)}$ are fixed at zero, and otherwise $\Phi^{(1)}$ is constrained to equal $\Phi^{(2)}$. In addition, in Model M four of the error variances were constrained to be equal between the two groups while the other four were allowed to differ. This last model is the closest yet to achieving acceptability. With $\chi^2\ (53\ df) = 73.28$, it falls just short of the $p > .05$ criterion for moving into the category of good models.

Table 7 presents a shortened and concise history of the search for the class of good models by listing five models, all of which are accepted by the χ^2 criterion. The last model, Model M in the intermediate phase just discussed, is the starting point for finding the first of the good models, Model N in Table 7. It will be remembered that in this model (Model M), the factor loadings were equal for the two groups and correlations among the factors were also equal for the two groups. The only difference between the groups was in four of the error variances.

Table 7: Final phase in the search for acceptable models. Five hypotheses about the model-implied Σs and the acceptability of the fit to the data.

Model	χ^2	df	p-value	Decision
N	60.05	51	.18	Accept
O	54.00	50	.32	Accept
P	48.33	49	.50	Accept
Q	48.23	48	.46	Accept
R	41.62	47	.69	Accept

However, something else needed to be changed in order to find a better fit. LISREL 8 possesses the convenient feature of pinpointing which parameter should be freed next in order to achieve the largest drop in χ^2. In Model M above, it wanted the covariance between the errors of two tests to be freed to accomplish this objective. Normally, we assume at first that the errors among the tests are uncorrelated, but, in this case, to reach the class of good models we have to allow a correlation between the errors on MCT and SAT for both groups. This means that there is some degree of association between the specific factors for these two tests. Once we allow this correlation between the errors on these two tests for the two groups, we have a significant drop in the χ^2 from 73.28 to 60.05 as exhibited by Model N in the first row of Table 7. We lose only 2 df for the two new estimates with the result that the p value climbs above the .05 criterion level and gives us our first acceptable model.

The remaining models assume everything in Model N and progressively relax an additional parameter. These are the parameters LISREL points to as causing the best drop in χ^2. Model O frees up $\lambda_{52}^{(2)}$, the factor loading of MVTP on ξ_2 for the computer group. Model P frees up $\theta_{41}^{(1)}$, the correlation between errors for MVT and ANI for the paper and pencil group. Model Q frees up $\lambda_{53}^{(1)}$, the factor loading of MVTP on ξ_3 for the paper and pencil group. Finally, in Model R, the same factor loading from Model D is freed for the computer group. There is one df lost for each one of these new estimates.

Obviously, with each new model the interpretation becomes harder and harder. The degree of psychometric equivalency between the two groups is increasingly watered down with each new destruction of an equality constraint. Also, there is an increasing chance of attempting to explain noise in the data as opposed to the actual signal. For these reasons, it is best to stick with the simplest model that is acceptable. As mentioned before, no one model captures the absolute truth, but our tentative working model while we wait for further confirmation or disconfirmation from more data will be Model N in the class of good models.

The actual estimates produced by LISREL for the three parameter matrices under this tentative working model will be examined in detail in the next section. The important point is that a model with an acceptable fit to the sample covariance data has been found. This model exhibits fairly strong psychometric equivalency between the two groups. Explicitly, this statement means that following equalities listed below in Equations (17) through (23) were in place,

$$\mathbf{\Lambda_x^{(1)}} = \mathbf{\Lambda_x^{(2)}} \tag{17}$$

$$\mathbf{\Phi^{(1)}} = \mathbf{\Phi^{(2)}} \tag{18}$$

$$\theta_{22}^{(1)} = \theta_{22}^{(2)} \tag{19}$$

$$\theta_{33}^{(1)} = \theta_{33}^{(2)} \tag{20}$$

$$\theta_{66}^{(1)} = \theta_{66}^{(2)} \tag{21}$$

$$\theta_{77}^{(1)} = \theta_{77}^{(2)} \tag{22}$$

$$\theta_{32}^{(1)} = \theta_{32}^{(2)} \tag{23}$$

The number of latent variables ($n = 4$) are the same for the two groups and match up with the test content. The factor loadings of the eight tests on the four factors assume an especially simple form for both groups, and the correlations between the factors are the same for both groups. The independence between the underlying Spatial Apperception skill (ξ_3) and the other three skills (ξ_1, ξ_2 and ξ_4) seems quite reasonable. The only fly in the ointment is the association between MCT and SAT ($\theta_{32}^{(1)} = \theta_{32}^{(2)}$) not accounted for by their respective factors. In addition, four of the eight error variances are equal. To achieve a good fit we had to allow the error variances on four of the tests to be larger for the paper and pencil group than for the computer group. Had this been in the other direction there would have been cause for concern. The computer mode of presentation, however, seems to be the same or better in terms of test reliability when compared to the paper and pencil presentation format.

Examination of LISREL Parameter Estimates for the Good Model

We now examine the actual estimates and uncertainties of these estimates for all the elements of the three parameter matrices. These estimates were all made by the maximum likelihood method. Within the LISREL program, $\mathbf{\Lambda_x^{(1)}}$ was constrained to equal $\mathbf{\Lambda_x^{(2)}}$ for Model N of Table 7. This led to a non-significant χ^2, so such a model provided an acceptable fit to the two sample variance covariance matrices. See Table 8 for the common factor loading matrix, $\mathbf{\Lambda_x}$, for the two groups. These factor loadings are not estimates by the LISREL program. They were fixed beforehand by the model that tied the tests as outward indicators to the corresponding latent variables.

Table 8: The loadings of each of the eight tests on the four underlying latent variables for both groups. This is the $\Lambda_x^{(1)} = \Lambda_x^{(1)}$ equality constraint of Model N in the class of good models.

Test	ξ_1	ξ_2	ξ_3	ξ_4
MVT	1	0	0	0
MCT	0	1	0	0
SAT	0	0	1	0
ANI	0	0	0	1
MVTP	1	0	0	0
MCTP	0	1	0	0
SATP	0	0	1	0
ANIP	0	0	0	1

As part of that model, $\Phi^{(1)}$ was constrained to equal $\Phi^{(2)}$ in addition to the equality of the factor loadings. The factor structure and factor loadings, as well as the correlations between the factors, are the same for the two groups. Table 9 gives the maximum likelihood estimates for Φ as provided by LISREL. The standard error and the t-value for assessing whether the given estimate is significantly different than zero are shown below the estimate. From theoretical considerations, the Spatial Apperception skill (ξ_3) was judged independent of the other three skills (ξ_1, ξ_2, and ξ_4). This explains the 0 values with no associated standard error that appear in the table. The estimate of the covariance between the Math/Verbal skill (ξ_1) and Aviation/Nautical Interest (ξ_4) is not significant. These two factors may also be uncorrelated.

Table 9: The estimates of the variances and covariances among the four underlying latent variables for Model N in the class of good models. This is the common matrix for both paper and pencil and computer presentation groups because $\Phi^{(1)}$ was specified as equal to $\Phi^{(2)}$ in the LISREL program.

	ξ_1	ξ_2	ξ_3	ξ_4
ξ_1	18.25 (3.51) 5.19			
ξ_2	8.13 (1.93) 4.21	7.42 (1.78) 4.17		
ξ_3	0.00	0.00	21.28 (4.02) 5.29	
ξ_4	1.92 (1.70) 1.14	5.81 (1.36) 4.28	0.00	7.78 (1.71) 4.56

It is very informative to re-cast these estimates of the covariances among the factors into correlation coefficients. By definition, the entries in the off-diagonal positions of any Σ matrix are $\rho_{ij}\sigma_i\sigma_j$, where ρ_{ij} is the correlation coefficient between the ith and jth factors, and σ_i is the standard deviation of the ith factor while σ_j

is the standard deviation of the jth factor. For example, the estimate of the correlation between Quantitative/Verbal Skill (ξ_1) and Mechanical Comprehension Skill (ξ_2) is,

$$\begin{aligned} \rho_{12}\sigma_1\sigma_2 &= 8.13 \\ \rho_{12}\sqrt{18.25}\sqrt{7.42} &= 8.13 \\ \rho_{12} &= \frac{8.13}{11.637} \\ &= .70 \end{aligned}$$

where we have substituted the estimates of σ_1 and σ_2. The other two correlation coefficients which need to be computed are shown in Table 10 below. Both correlations between Quantitative/Verbal Skill and Mechanical

Table 10: The estimates of the correlation coefficients among the four underlying latent variables for Model N in the class of good models.

	ξ_1	ξ_2	ξ_3	ξ_4
ξ_1	1.00			
ξ_2	.70	1.00		
ξ_3	.00	.00	1.00	
ξ_4	.16	.76	.00	1.00

Comprehension Skill and between Mechanical Comprehension Skill and Aviation/Nautical Interest are rather large, while, as mentioned above, the small correlation between Quantitative/Verbal Skill and Aviation/Nautical Interest may not be significantly different from 0. Since the Spatial Apperception Skill was judged from a theoretical basis to be independent of the other three skills, it has a zero correlation with these other underlying factors. These correlation coefficients among the latent variables are called "disattentuated correlations" because they have been freed from the smaller (attenuated) correlations that would have been surmised from considering just the fallible tests.

Both Λ_x and Φ can be set equal for the two modes of presentation, so the parameter matrix where the differences between the two groups occurs has been isolated. The deviation from psychometric equality must reside in Θ_δ which contain the variances of the specific factor and the measurement error for each of the eight tests. Although Θ_δ is an 8×8 symmetric matrix, we need show only the eight variances that occur along the diagonal of Θ_δ. The one exception to this statement is the covariance between MCT and SAT which is needed to reach the class of good models. Table 11 gives the LISREL estimates of the diagonal elements of Θ_δ as θ_{ii} for both groups. The standard error and t-values are given for these estimates as in the previous tables. Even here, the error variances for δ can be constrained to be equal for four of the tests and an acceptable χ^2 can still be achieved. As can be seen from Table 11, the tests MCT, SAT, MCTP, and SATP are constrained to have equal variances for δ.

So the reason that an extremely strong form of psychometric equivalency cannot be adopted is ascribed to differences in test reliability in four of the tests. However, for each of these tests the computer group had the lower error variance, or what amounts to the same thing, the higher test reliability. The surprising and encouraging bottom line of this analysis is that presenting the ASTB via the computer causes no disruption whatsoever to the

Table 11: The estimated variance for the specific factor and measurement error of the eight tests for both groups. Four of the tests were constrained to have equal values between the two groups.

	MVT θ_{11}	MCT θ_{22}	SAT θ_{33}	ANI θ_{44}	MVTP θ_{55}	MCTP θ_{66}	SATP θ_{77}	ANIP θ_{88}
Paper and Pencil	12.46	7.86	9.58	11.20	10.91	7.08	8.49	8.28
	(4.12)	(1.58)	(2.80)	(3.13)	(3.87)	(1.45)	(2.58)	(2.58)
	3.03	4.96	3.42	3.58	2.82	4.87	3.29	3.21
Computer	7.39	7.86	9.58	3.68	1.99	7.08	8.49	2.54
	(2.42)	(1.58)	(2.80)	(1.27)	(1.78)	(1.45)	(2.58)	(1.11)
	3.05	4.96	3.42	2.89	1.12	4.87	3.29	2.28

underlying psychometric properties *except that some of the tests presented by the computer have less specific and measurement error attached.* The most likely explanation is that, simply because of sampling differences, the individuals in the computer group had less specific error on the MVT and ANI. This was true as well on these subtests the previous time they took them. The mode of presentation did not impact this pattern of the unaccountable causes all lumped into this one catch-all factor.

Testing Equality of the Means

Changing the mode of presentation for the ASTB apparently did not adversely affect the variances and covariances. But, by definition, the preceding analysis of the covariances ignored any potential effects of the mode of test presentation on the means of the two groups. LISREL is also capable of determining whether changing the mode of presentation from a paper and pencil format to a computer format impacted the means of the test scores in any significant way.

Two new parameter matrices (actually vectors) are introduced by LISREL for this discussion of the means. They are τ_x and κ. Equation (1) is amended to,

$$x = \tau_x + \Lambda_x \xi + \delta \tag{24}$$

so that τ_x is seen to be an intercept term, while κ is defined as the mean of ξ. The means of the four latent variables ξ_1 through ξ_4, which together constitute ξ for the paper and pencil group, are set at,

$$\kappa^{(1)} = (0, 0, 0, 0) \tag{25}$$

The means of ξ for the computer group will be estimated by LISREL and placed in $\kappa^{(2)}$. A t-test will be used to determine whether any of the estimates in $\kappa^{(2)}$ are significantly different from zero.

For the computer group, $\kappa^{(2)}$ was estimated as,

$$\kappa^{(2)} = (.55, -1.03, .50, 1.02) \tag{26}$$

See Table 12 for the standard errors and t values of these estimates. None of the t-values are significant, so we can conclude that the means of the four underlying latent variables for the computer group do not differ from the means of the corresponding underlying latent variables for the paper and pencil group. Just as in the previous

Table 12: The maximum likelihood estimates for the means, $\kappa^{(2)}$, of the four latent variables of the computer group. They do not differ from 0 and therefore do not differ from $\kappa^{(1)}$, the means of the paper and pencil group.

	ξ_1	ξ_2	ξ_3	ξ_4
Estimate	0.55	−1.03	0.50	1.02
Standard Error	(1.05)	(0.74)	(1.13)	(0.74)
t-value	0.52	−1.39	0.44	1.39

analysis of the covariances, there seems to be no adverse impact on the test score means when the ASTB is presented in a computerized format.

The equality constraints that define psychometric equivalency for Λ_x, Φ, and Θ_δ were the same as Model N in the last section. There exists an additional constraint for the model of the means, Equation (24), that includes the intercept term. This additional constraint is,

$$\tau_x^{(1)} = \tau_x^{(2)} \tag{27}$$

The chi-square for this model with all of these constraints was,

$$\chi^2\,(55\ df) = 67.93 \qquad p = .11 \tag{28}$$

This particular model belongs to the class of acceptable models. Table 13 provides the common estimates of τ_x for both groups. Since $\kappa^{(1)} = \kappa^{(2)} = 0$, the intercept term, τ_x, then is very close to the observed means of the

Table 13: The LISREL maximum likelihood estimates for the intercept term, τ_x, in the measurement model of Equation (24).

	MVT τ_1	MCT τ_2	SAT τ_3	ANI τ_4	MVTP τ_5	MCTP τ_6	SATP τ_7	ANIP τ_8
Estimate	27.14	22.51	26.38	18.89	27.41	22.73	26.70	18.32
Standard Error	(0.80)	(0.56)	(0.83)	(0.59)	(0.79)	(0.56)	(0.82)	(0.57)
t-value	34.01	40.33	31.97	32.05	34.65	40.74	32.40	31.91

actual scores on the eight tests as averaged over the two groups. As is to be expected, these are all highly significantly different from 0.

Equivalency for Criterion Measures

At this point in the report we have examined some internal psychometric properties for two modes of test presentation. That is, we have looked at the first two objectives set out in the Introduction. The equality constraints placed on the three parameter matrices, Λ_x, Φ, and Θ_δ answered the first objective concerning the psychometric properties reflected in the variances and covariances of the eight tests. Next, we examined τ_x and κ to answer the second objective concerning any possible differences in means caused by mode of presentation. Now

it is time to address the third concern, the question of validity of the two presentation modes with some external measure of flight training success.

The LISREL measurement model for x was, by itself, sufficient to address the first two objectives. The analysis must now, however, be expanded to include the structural equation part of the LISREL model in order to deal with predictive validity.

$$y = \Lambda_y \eta + \varepsilon \tag{29}$$

$$\eta = \Gamma \xi + \zeta \tag{30}$$

$$y = \Lambda_y (\Gamma \xi + \zeta) + \varepsilon \tag{31}$$

Equation (29) is the measurement model for the criterion variables contained in the column vector y. It is completely analogous to the measurement model for the test scores used as predictor variables in x. η is the vector of latent variables for the criterion and ε contains the measurement errors for y. The measurement model for y is simplified greatly because we take η to be the same as the criterion score itself. Therefore, the factor loadings are 1 in Λ_y and Θ_ε is zero. Equation (30) is the structural equation part of the LISREL model that captures the putative causal relationship connecting the predictor latent variables to the latent variables of the criterion measure. Γ is the parameter matrix that contains the structural coefficients in this causal regression-like relationship. When the structural equation of Equation (30) is substituted into the measurement equation of Equation (29), Equation (31) results.

Does the presentation of test items via the computer result in a different set of structural coefficients for some criterion variables as compared to the traditional presentation via paper and pencil? Our main concern is with estimating the parameter elements in Γ for the two groups to see if they can be considered statistically equivalent. If so, then we have further confirmation for the psychometric equivalency of the two modes of presentation.

The initial analysis of the training criteria data is very limited. It is circumscribed by the lag time between the gathering of the test scores in the laboratory and the slow maturation of the criterion data as the subjects complete what may be a two to three year flight training program. We hope to present a more in-depth analysis at a later date. Therefore, the analysis contained in this section of the report is more of an example of a preliminary feasibility check of what can be done with LISREL.

The one criterion measure for which complete data is available is employed in this present analysis. This is the composite score on overall performance in Aviation Preflight Instruction (API), the ground school portion of flight training before students enter the actual flying curriculum in Primary Training. Of the 82 subjects in the experiment, 73 subjects successfully completed training through Primary. From long standing historical data we know that about 10% of the students have attrited after Primary. The 9 attritions from a total of 82 fall perfectly in line with this attrition rate.

Of the 9 subjects who attrited, 3 subjects attrited in API and 6 subjects attrited in Primary. No scores were given by the training command for the 2 of the 3 attritions in API. Therefore, we have to assume that they fell below some threshold on a normal distribution. A criterion score was assigned to these two attritions by setting the threshold at -2σ for the normal distribution defined by $\mu = 50$ and $\sigma = 10$. This mean and standard deviation define the Navy Standard Score for API performance. Therefore, the threshold exists at a API criterion score of 30. The two attritions were randomly given criterion scores of 30 and 29. The one attrition given a score by the

training command was 32. After this assignment, the mean for the criterion score over the 82 subjects was 50.60 with a SD of 7.41.

Because there is only one criterion variable, only one variable appears then in the vector $\boldsymbol{y}$. Equation (31) above becomes

$$y = \Gamma\xi + \zeta \tag{32}$$

when we make the assumption mentioned previously that the latent variable is the same as the criterion score itself. With only one criterion variable to worry about, Γ becomes a 1×4 row vector with individual elements, $\gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{14}$. Equation (32) can then be translated from vector and matrix notation into a scalar equation as

$$y = \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \gamma_{13}\xi_3 + \gamma_{14}\xi_4 + \zeta \tag{33}$$

Can an acceptable fit be found for the new augmented variance-covariance matrix that includes the criterion variable when the Γ parameters are constrained to be equal in the two groups? All other constraints are as they existed in our tentative working Model N of the good model category. With the addition of the criterion variable, the overall number of different elements in the two sample data matrices now numbers 90 as compared to the previous number of 72. Therefore, there are 90 df for this problem.

The goodness of fit of this model to the data with the Γ parameters and the one ψ parameter equal for the two groups is

$$\chi^2(64\ df) = 77.41 \qquad p = .12 \tag{34}$$

We cannot reject this model's fit to the data, so the hypothesis that the computer mode of presentation does not affect test score association to a criterion measure is a tenable one.

It is instructive to pinpoint exactly where 26 degrees of freedom are used up so that the χ^2 test has only 64 df. Table 14 lists the parameter matrices and the number of elements that were estimated for each matrix. One df was lost for each estimate so this accounts for the 64 df in the χ^2 test.

Table 14: The various parameter matrices in the LISREL structural equation model showing how many estimates need to be made after any equality constraints have been taken care of.

Parameter Matrix	Number of estimates
Γ	4
Φ	7
$\Theta_\delta^{(1)}$	9
$\Theta_\delta^{(2)}$	5
Ψ	1
Sum	26

The estimates, standard errors, and t-values of the estimates for both groups are contained in Table 15. The structural coefficient for ξ_3, γ_{13}, is observed to possess the only significant t-value. The Spatial Apperception skill is the only one of the four skills measured by the ASTB to be associated with overall success in API. None of

the other underlying skills, Quantitative/Verbal (ξ_1), Mechanical Comprehension (ξ_2), or Aviation/Nautical Interest (ξ_4) were significantly associated with the criterion score.

Table 15: The maximum likelihood estimates for the common structural equation parameters.

	γ_{11}	γ_{12}	γ_{13}	γ_{14}	ψ_1
Estimate	1.54	−2.10	0.66	1.83	29.75
Standard Error	(1.89)	(4.89)	(0.16)	(3.27)	(10.17)
t-value	0.82	−0.43	4.13	0.56	2.93

By selection testing standards, the structural equation exhibits a fairly healthy relationship to the criterion. The squared multiple correlation coefficient is $R^2 = .46$ for these data. The variance of the error term, ζ, in the structural equation, Equation (33), is contained in the matrix Ψ, which in this case is only one element, ψ_1. This estimate is listed in the final column of Table 15 and its relatively large size is the inhibiting reason why R^2 is not larger. The estimated SD of the normal distribution for ζ is about 5.45. Since the variance of the criterion score is estimated at 55.23 and the variance of the error is 29.75, R^2 can be interpreted as the fraction of the total variance explained by the structural equation.

$$R^2 = 1 - \frac{29.75}{55.23} \tag{35}$$

$$= .46 \tag{36}$$

We are interested in examining this pattern of results when other criterion variables from Primary Training are examined. However, to reiterate, the important point for this study is finding a acceptable model where the four elements of the vector containing the structural equation coefficients could be set equal for the two modes of presentation. As before, had we been forced to relax this equality constraint in order to find an acceptable model, it would have called into question the hypothesis of the negligible influence on test scores of a computerized format for the ASTB.

Summary

In order to bring the APEX system to a fully operational status, we have to understand the potential effects of the system to change subjects's scores on the ASTB. In this initial foray, we examined the impact of presenting the ASTB in a computerized format as compared to the traditional paper and pencil version. Psychometric properties of the tests were defined in terms of a structural equation model. The software program LISREL 8 was used to find several models that fit the observed data and to find the estimates of the parameters within any one model. The simplest acceptable model was retained as the tentative working hypothesis concerning the changes in the psychometric properties brought about by the two different ways of presenting the ASTB.

Acceptable models can be found that allow for a very strong sense of psychometric equivalency between the two modes of presentation. The number of factors, the loadings of the sub-tests on the factors, and the correlations among the factors can be constrained to be equal for the two groups. The only equality constraint that has to be

relaxed is the one governing the variances of the combined specific and measurement error for some of the tests. But, even in this case, the direction of the change was to *reduce* the error variance for the computer group.

This first analysis concentrated on the measurement model component of the overall structural equation model. Therefore, only the variance-covariance matrix was needed as data in this kind of internal reliability and validity analysis. However, it is possible that the APEX system keeps the correlational structure of the sub-tests intact, but greatly changes their means. This obviously would be a very important impact of the new system. Using the same strategy of postulating equality for the relevant parameters to see if an acceptable model could be achieved, equality of the means over the sub-tests for the two groups could not be rejected either. Apparently, this facet of the test taker's performance was also not adversely affected by the new system.

Finally, we took an early, preliminary look at one of the criterion variables for which we had complete data. The purpose here was to see if the mode of presentation affected the structural coefficients in a regression equation linking the ASTB subtests as predictor variables to the overall performance in ground school as the criterion variable. Again, when equality constraints were placed on the structural coefficients and the error term between the two groups, an acceptable model was found.

The results of this analysis were quite favorable to the view that the APEX system does not cause unwanted changes in the assessment of the test taker's underlying cognitive skills as defined by the ASTB. As a caveat, we must hasten to point out that this initial examination of the impact of the APEX system on psychometric equivalency is based on relatively small sample sizes. Therefore, the power of the experiment to detect any real changes attributable to the new system, should such changes actually exist, is compromised. In addition, it must be kept in mind that, within the class of acceptable models, there are many models that allow for a more complex interpretation of the psychometric equivalency issues. We have opted for *one* of these acceptable models that supports the case for a simple and strong equivalency. We continue to gather further data to probe the robustness of this tentative working hypothesis.

In conclusion, based on the data and analysis contained in this report, we retain the belief that the technological improvements offered by the APEX system are in no way offset by any unpleasant disruption to the psychometric properties of the ASTB.

References

1. Hayduk, L. A. *Structural Equation Modeling with LISREL.* The Johns Hopkins University Press, Baltimore, MD, 1987.

2. Loehlin, J. C. *Latent Variable Models: An introduction to factor, path, and structural analysis.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.

3. Bollen, K. A. *Structural Equations with Latent Variables.* John Wiley & Sons, New York, NY, 1989.

4. Jöreskog, K. G. and Sörbom, D. *LISREL 8: User's Reference Guide.* Scientific Software International, Chicago, IL, 1996.

5. Biggerstaff, S., Portman, C., Blower, D. and Chapman, A. The Effect of Presentation Medium on Pilot Selection Test Battery Scores. NAMRL Technical Report. Pensacola, Fl. Under review. 1997.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	23 April 1998	

4. TITLE AND SUBTITLE
Psychometric Equivalency Issues for the APEX System

5. FUNDING NUMBERS
62233N MM33-30.001-7602

6. AUTHOR(S)
D. J. Blower

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Aerospace Medical Research Laboratory
51 Hovey Road
Pensacola Fl 32508-1046

8. PERFORMING ORGANIZATION REPORT NUMBER
NAMRL Special Report 98-1

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Naval Medical Research and Development Command
National Naval Medical Center
Building 1, Tower 12
8901 Wisconsin Avenue
Bethesda, MD 20889-5606

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This report discusses some of the psychometric properties of the Aviation Selection Test Battery (ASTB) when administered in the standard paper and pencil format as compared to an experimental version administered on the computer. Structural Equation Modeling is employed as the analytical framework to address certain questions of psychometric equivalency. The two experimental conditions are compared in their effects on the internal reliability and validity on eight subtests of the ASTB, the means of these scores, and finally on the structural coefficients when these scores are used in a predictive validity setting. The two modes of presentation for the ASTB, the traditional paper and pencil version and the experimental computer version, do not seem to differ in many psychometric properties. These results, then, are important in deflecting criticism concerning the impact of this new presentation medium on ASTB test scores.

14. SUBJECT TERMS
Aviation selection, Structural equation modeling, LISREL, Psychometric equivalency, APEX system

15. NUMBER OF PAGES
21

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR